Sur l'estimation du tau de Kendall conditionnel à l'aide de méthodes de classification

Alexis Derumigny [1] & Jean-David Fermanian [2]

[1] *CREST-ENSAE, 5 avenue Henry Le Chatelier, 91764 Palaiseau Cedex, France.*
alexis.derumigny@ensae.fr
[2] *CREST-ENSAE, 5 avenue Henry Le Chatelier, 91764 Palaiseau Cedex, France.*
jean-david.fermanian@ensae.fr. *This research has been supported by the Labex Ecodec.*

Résumé. Nous montrons ici comment le problème d'estimation du tau de Kendall conditionnel peut être réécrit comme un problème de classification. Le tau de Kendall conditionnel est un paramètre de dépendance conditionnel qui est caractéristique d'une paire de variables aléatoires. Le but est de prédire si la paire est concordante (valeur de 1) ou discordante (valeur de -1) conditionnellement à un vecteur de covariables. Nous prouvons la consistance et la normalité asymptotique d'une famille d'estimateurs basés sur un maximum de vraisemblance approché, comportant comme cas particuliers l'équivalent des régressions logit et probit dans notre cadre. Nous détaillons des algorithmes spécifiques, adaptant les techniques usuelles d'apprentissage automatique comme les plus proches voisins, les arbres de décision, les forêts aléatoires et les réseaux de neurones dans le contexte de l'estimation du tau de Kendall conditionnel. Des simulations détaillent leurs propriétés à distance finie. Finalement, tous ces estimateurs sont appliqués à une base de données constituée d'indices boursiers européens.

Mots-clés. Tau de Kendall conditionnel, mesure de dépendance conditionnelle, apprentissage automatique, problème de classification, indices boursiers

Abstract. It is shown how the problem of estimating conditional Kendall's tau can be rewritten as a classification task. Conditional Kendall's tau is a conditional dependence parameter that is a characteristic of a given pair of random variables. The goal is to predict whether the pair is concordant (value of 1) or discordant (value of -1) conditionally on some covariates. The consistency and the asymptotic normality of a family of penalized approximate maximum likelihood estimators is proven, including the equivalent of the logit and probit regressions in our framework. Specific algorithms are detailed, adapting usual machine learning techniques, including nearest neighbors, decision trees, random forests and neural networks, to the setting of the estimation of conditional Kendall's tau. Finite sample properties of these estimators are assessed in a simulation study. Finally, all these estimators are applied to a dataset of European stock indices.

Keywords. conditional Kendall's tau, conditional dependence measure, machine learning, classification task, stock indices.

1 Introduction

Mis à part la corrélation linéaire, la plupart des mesures de dépendance entre deux variables aléatoires sont des fonctions de la copule sous-jacente uniquement : le rho de Spearman, le tau de Kendall, le beta de Blomqvist, le coefficient de Gini, etc. En conséquence, ils sont indépendants des marges correspondantes, ce qui est considéré comme un point positif. Quand des covariables sont disponibles, ces mesures admettent une extension naturelle, définissant ainsi des mesures de dépendance "conditionnelles". En théorie, il suffit de remplacer la copule par une copule conditionnelle pour obtenir directement une "version conditionnelle" de n'importe quelle mesure de dépendance. De telles mesures ont été peu étudiées jusqu'à présent, après leur introduction par Gijbels, Veraverbeke, et Omelka (2011a, 2011b).

Maintenant, nous introduisons notre mesure de dépendance clé : pour un entier p fixé, et pour chaque $\mathbf{z} \in \mathbb{R}^p$, le tau de Kendall conditionnel d'un vecteur bivarié $\mathbf{X} := (X_1, X_2)$ sachant un vecteur de covariables $\mathbf{Z} = \mathbf{z}$ est défini par

$$\begin{aligned}\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) \\ - \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) < 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big),\end{aligned}$$

où $(\mathbf{X}_1, \mathbf{Z}_1) = (X_{1,1}, X_{1,2}, Z_{1,1}, \ldots, Z_{1,p})$ et $(\mathbf{X}_2, \mathbf{Z}_2) = (X_{2,1}, X_{2,2}, Z_{2,1}, \ldots, Z_{2,p})$ sont deux copies indépendantes de $(\mathbf{X}, \mathbf{Z})$. Nous supposons que la loi de $\mathbf{X}$ sachant $\mathbf{Z} = \mathbf{z}$ est absolument continue par rapport à la mesure de Lebesgue, et ce pour chaque valeur $\mathbf{z}$. Ceci implique que

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = 2\,\mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - 1.$$

On peut noter que le tau de Kendall conditionnel appartient toujours à l'intervalle $[-1, 1]$ et reflète une dépendance positive (si $\tau_{1,2|\mathbf{Z}=\mathbf{z}} > 0$) ou négative (si $\tau_{1,2|\mathbf{Z}=\mathbf{z}} < 0$) entre X_1 et X_2, sachant $\mathbf{Z} = \mathbf{z}$. Contrairement aux corrélations, cette mesure a l'avantage d'être toujours bien définie, même si l'un des X_k, $k = 1, 2$, n'admet pas de moments d'ordre 2. C'est le cas lorsqu'il suit une loi de Cauchy par exemple.

Plusieurs estimateurs du tau de Kendall conditionnel ont déjà été proposés dans la littérature, soit comme sous-produit de l'estimation de copules conditionnelle par Gijbels, Veraverbeke, et Omelka (2011a) et par Fermanian et Lopez (2018) - ou directement, par Derumigny et Fermanian (2018a, 2018b). Néanmoins, il ne nous semble pas que quelqu'un ait remarqué la relation entre le tau de Kendall conditionnel et les méthodes de classification.

Cette relation est la suivante. Soient $W := 2 \times \mathbb{1}\{(X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0\} - 1$ et $\mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) = \mathbb{P}\big(W = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) =: p(\mathbf{z})$. Nous pouvons alors remarquer que la prédiction de la concordance/discordance parmi les paires d'observations $(\mathbf{X}_1, \mathbf{X}_2)$ sachant $\mathbf{Z}$ peut être vue comme un problème de classification de

ces paires. Si un modèle est capable d'estimer la probabilité conditionnelle d'observer des paires concordantes d'observations, alors il est capable d'estimer le tau de Kendall conditionnel. De telles probabilités conditionnelles sont justement les sorties données par la plupart des algorithmes de classification usuels. Ainsi, la plupart des classifieurs peuvent être utilisés ici (par exemple les classifieurs linéaires, les arbres de décision, les forêts aléatoires, les réseaux de neurones, et ainsi de suite), mais appliqués ici à des paires d'observations.

Autrement dit, pour chaque $1 \leq i, j \leq n$, $i \neq j$, soit $W_{(i,j)}$ défini par

$$W_{(i,j)} := 2 \times \mathbb{1}\{(X_{j,1} - X_{i,1})(X_{j,2} - X_{i,2}) > 0\} - 1 = \begin{cases} 1 & \text{si } (i,j) \text{ est concordante,} \\ -1 & \text{si } (i,j) \text{ est discordante.} \end{cases} \quad (1)$$

Un algorithme de classification va répartir un couple donné (i, j) dans l'une des deux catégories $\{1, -1\}$ (ou "concordant versus discordant", de façon équivalente), avec une certaine probabilité, sachant la valeur de la variable explicative commune $\mathbf{Z}$.

2 Utilisation d'un modèle de type régression

Un premier modèle simple est celui dans lequel

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = g(\boldsymbol{\psi}(\mathbf{z})^T \beta^*), \quad \forall \mathbf{z} \in \mathcal{Z}, \quad (2)$$

où $\boldsymbol{\psi} : \mathbb{R}^p \to \mathbb{R}^{p'}$ et g sont connus, pour des entiers $p, p' > 0$ et un ensemble $\mathcal{Z} \subset \mathbb{R}^p$, en suivant l'article de Derumigny et Fermanian (2018a). Ici, nous proposons une procédure d'inférence de β^* directe, contrairement à Derumigny et Fermanian (2018a) où une procédure en deux étapes était nécessaire.

Une des difficultés de l'estimation de ce modèle réside dans les évènements conditionnants, qui sont de la forme $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$ et nécessiteront l'utilisation de techniques de localisations. Plus précisément, on considérera comme "pertinents" les couples d'observations $\mathbf{X}_i$ et $\mathbf{X}_j$ pour lesquels les variables explicatives $\mathbf{Z}_i$ et $\mathbf{Z}_j$ correspondantes sont proches d'une valeur donnée $\mathbf{z}$. Si les variables $\mathbf{Z}$ étaient discrètes, on considérerait un sous-ensembles constitués des couples pour lesquels $\mathbf{Z}_i = \mathbf{Z}_j$. Dans notre cas où les $\mathbf{Z}$ sont continues, cet évènement ne se produit presque sûrement jamais, et il faut donc utiliser des techniques de lissage.

Soit K un noyau de dimension p, $h > 0$ une fenêtre de lissage, et $K_h(\mathbf{z}) := K(\mathbf{z}/h)/h^p$. La log-vraisemblance associée à l'observation $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ sachant $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$ est

$$\begin{aligned} \ell_\beta(W_{(i,j)}, \mathbf{z}) := & \left(\frac{1 + W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\Big(W_{(i,j)} = 1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\Big) \\ & + \left(\frac{1 - W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\Big(W_{(i,j)} = -1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\Big). \end{aligned}$$

En pratique, comme la loi de $\mathbf{Z}$ est continue, il n'y a presque sûrement aucun couple pour lequel $\mathbf{Z}_i = \mathbf{Z}_j$. En conséquence, on utilisera une log-vraisemblance approchée, basée sur l'échantillon $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ pour toutes les paires (i, j), $i \neq j$. Cette log-vraisemblance approchée est définie par

$$L_n(\beta) := \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j)\ell_\beta(W_{(i,j)}, \tilde{\mathbf{Z}}_{i,j}),$$

pour un choix de $\tilde{\mathbf{Z}}_{i,j}$ dans un voisinage de $\mathbf{Z}_i$ ou de $\mathbf{Z}_j$. Par exemple, il peut être choisi appartenant à $\{\mathbf{Z}_i, \mathbf{Z}_j, (\mathbf{Z}_i + \mathbf{Z}_j)/2\}$. Ici, nous proposons

$$\begin{aligned} L_n(\beta) &:= \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j)\ell_\beta(W_{(i,j)}, \mathbf{Z}_i) \\ &= \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \left\{ \left(\frac{1 + W_{(i,j)}}{2} \right) \log \left(\frac{1}{2} + \frac{1}{2} g(\boldsymbol{\psi}(\mathbf{Z}_i)^T \beta) \right) \right. \\ &+ \left. \left(\frac{1 - W_{(i,j)}}{2} \right) \log \left(\frac{1}{2} - \frac{1}{2} g(\boldsymbol{\psi}(\mathbf{Z}_i)^T \beta) \right) \right\}. \end{aligned}$$

Sous (2), on peut en déduire un estimateur de β^* basé sur la maximisation de cette fonction, avec une pénalité ℓ_1 (de type Lasso), c'est-à-dire,

$$\hat{\beta} := \arg\max_{\beta \in \mathbb{R}^{p'}} L_n(\beta) - \lambda_n |\beta|_1, \tag{3}$$

où λ_n est un paramètre à choisir. Sous des hypothèses de régularité détaillées dans l'article de Derumigny et Fermanian (2019), on obtient les résultats suivants.

Théorème 1. *Si $\lambda_n \to \lambda_\infty$ et $n^2 h^p \to \infty$ quand $n \to \infty$, et si $\beta \mapsto L_n(\beta)$ est concave (ce qui est le cas pour les g correspondants aux probit et logit), alors la solution $\hat{\beta}$ de (3) tend en probabilité vers $\beta^{**} := \arg\max_\beta L_\infty(\beta) - \lambda_\infty |\beta|_1$, où*

$$L_\infty(\beta) := \int \phi(\mathbf{z}, \mathbf{z}, \beta) f_{\mathbf{Z}}^2(\mathbf{z})\, d\mathbf{z},$$

$\phi(\mathbf{x}, \mathbf{y}, \beta) := p(\mathbf{x}, \mathbf{y}) \log(q(\mathbf{x}, \beta)) + (1 - p(\mathbf{x}, \mathbf{y})) \log(1 - q(\mathbf{x}, \beta))$, et $p(\mathbf{x}, \mathbf{y}) := \mathbb{P}_{\beta^*}((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 | \mathbf{Z}_1 = \mathbf{x}, \mathbf{Z}_2 = \mathbf{y})$.

Théorème 2. *Si $n^{1/2}\lambda_n \to \mu$ et $nh^p \to \infty$ quand $n \to \infty$, alors $n^{1/2}(\hat{\beta} - \beta^*)$ tend faiblement vers*

$$\mathbf{u}^* = \arg\min_{\mathbf{u} \in \mathbb{R}^p} \mathbb{W}(\beta^*)\mathbf{u} + \frac{1}{2}\mathbf{u}^T \mathbb{H}(\beta^*)\mathbf{u} - \mu \sum_{k;\beta_k^*=0} |u_k| - \mu \sum_{k;\beta_k^*\neq 0} sign(\beta_k^*) u_k,$$

où $\mathbb{W}(\beta^*) \sim \mathcal{N}(0_p, \Sigma_{\beta^*})$, $\Sigma_{\beta^*} = \int \partial_\beta \phi(\mathbf{z}, \mathbf{z}, \beta^*) \partial_\beta \phi(\mathbf{z}, \mathbf{z}, \beta^*)^T f_{\mathbf{Z}}^3(\mathbf{z})\, d\mathbf{z}$ et

$$\mathbb{H}(\beta^*) = \int \partial^2_{\beta,\beta^T} \phi(\mathbf{z}, \mathbf{z}, \beta^*) f_{\mathbf{Z}}^2(\mathbf{z})\, d\mathbf{z}.$$

3 Utilisations d'algorithmes généraux de classification

Dans la section précédente, nous avons étudié une procédure de maximum de vraisemblance approchée pour estimer le tau de Kendall conditionnel, mais elle nécessite la connaissance préalable des fonctions $\boldsymbol{\psi}$ et g. En pratique, il est bien difficile de deviner ou de choisir de telles formes fonctionnelles, en particulier pour g. Pour améliorer l'estimation de nos tau de Kendall conditionnel, on peut utiliser le lien précédemment obtenu entre son estimation, et l'estimation de $\mathbb{P}\big(W_{(1,2)} = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big)$, qui est la probabilité $p(\mathbf{z})$ de classifier le couple $(1, 2)$ dans l'une des deux catégories (concordante ou discordante) sachant une valeur commune de leurs covariables $\mathbf{z}$. Formellement, la réponse à une telle question peut être directement obtenue par un algorithme de classification. Ainsi, au lieu d'estimer un modèle paramétrique donné, un algorithme de classification estimera "automatiquement" $p(\mathbf{z})$ par un certain $\hat{p}(\mathbf{z})$. Finalement, on estimera directement le tau de Kendall conditionnel par $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := 2\hat{p}(\mathbf{z}) - 1$.

Pour cela, la première étape est de transformer l'*échantillon initial* $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in (\mathbb{R}^{2+p})^n$, en un objet $\tilde{\mathcal{D}}$, appelé l'*échantillon de paires* (cf. Algorithme 1). Chaque élément de cet échantillon de paires est indexé par un entier $k \in \{1, \dots, n(n-1)/2\}$, qui correspond à une paire (non ordonnée) (i, j), $i \neq j$, d'observations de l'échantillon initial. Chaque observation de l'échantillon de paires est composé de 3 éléments : la covariable moyenne $\tilde{\mathbf{Z}}_k$, l'indicateur de concordance $W_k \in \{-1, 1\}$ et un poids V_k, correspondant à la pertinence de la paire. En effet, plus $\mathbf{Z}_i$ et $\mathbf{Z}_j$ sont proches, plus la paire est pertinente pour notre problème de classification.

Algorithme 1 : Algorithme pour créer l'échantillon de paires à partir de l'échantillon initial.

Entrées : L'échantillon initial $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
$k \leftarrow 0$;
pour $i \leftarrow 1$ **à** $(n-1)$ **faire**
  pour $j \leftarrow (i+1)$ **à** n **faire**
    $\tilde{\mathbf{Z}}_k \leftarrow (\mathbf{Z}_i + \mathbf{Z}_j)/2$;
    $W_k \leftarrow W_{(i,j)}$ comme défini dans l'équation (1) ;
    $V_k \leftarrow K_h(\mathbf{Z}_i - \mathbf{Z}_j)$;
    $k \leftarrow k + 1$;
  fin
fin
Soit $\mathcal{K} := \{k : V_k > 0\}$;
Sorties : L'échantillon de paires $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k \in \mathcal{K}}$.

Soit un algorithme de classification qui prend en paramètre un échantillon composé d'une variable binaire à expliquer, d'un vecteur de variables explicatives et d'un poids

pour chaque observation. On remarque que cette description correspond exactement aux caractéristiques de l'échantillon $\tilde{\mathcal{D}}$ construit ci-dessus. En conséquence, il suffit d'appliquer cet algorithme à l'échantillon $\tilde{\mathcal{D}}$. Le résultat d'un tel algorithme sera une fonction $\mathbf{z} \mapsto \hat{p}(\mathbf{z})$ qui associe à chaque point $\mathbf{z}$ la probabilité estimée d'observer une paire concordante conditionnellement à $\mathbf{z}$. On en déduit un estimateur du tau de Kendall conditionnel par la formule $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} = 2\hat{p}(z) - 1$.

4 Application à des données de marchés financiers

Maintenant, nous appliquons ces différents algorithmes à une base de données d'indices boursiers. On s'intéresse à la dépendance conditionnelle entre les rendements des indices MSCI France et MSCI Allemagne durant la période (26/08/2012 - 02/03/2018). La variable conditionnante est la volatilité journalière de l'Eurostoxx définie comme le ratio $(Max - Min)/F$, où F est la valeur à la fermeture.

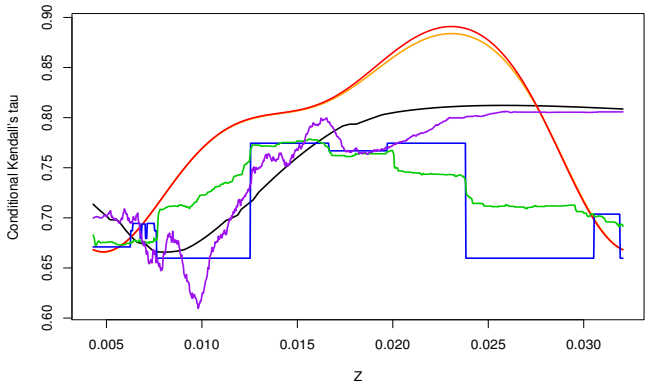

FIGURE 1 – Tau de Kendall conditionnel estimé par différents algorithmes de classification : logit (courbe orange), probit (courbe rouge), arbres de décision (courbe bleue), forêts aléatoires (courbe verte), plus proches voisins (courbe violette), réseaux de neurones (courbe noire).

Bibliographie

Derumigny, A. et Fermanian, J. D. (2018a), About Kendall's regression, *ArXiv preprint*, arXiv :1802.07613.

Derumigny, A. et Fermanian, J. D. (2018b), About kernel-based estimation of the conditional Kendall's tau : finite-distance bounds and asymptotic behavior, *ArXiv preprint*, arXiv :1810.06234

Derumigny, A., et Fermanian, J. D. (2019), A classification point-of-view about conditional Kendall's tau, *Computational Statistics & Data Analysis*.

Fermanian, J.-D. et Lopez, O. (2018), Single-index copulas, *Journal of Multivariate Analysis*, 165, 27–55.

Gijbels, I., Veraverbeke, N. et Omelka, M.(2011a), Conditional copulas, association measures and their applications, *Computational Statistics & Data Analysis*, 55 (5), 1919–1932.

Gijbels, I., Veraverbeke, N. et Omelka, M.(2011b), Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics* 38, 766–780